A Word Similarity Feature-based Semi-supervised Approach for Named Entity Recognition

Ze Wang
School of Control Science and Engineering
Dalian University of Technology
Dalian, China
wangze@mail.dlut.edu.cn

Zhongyang Han
School of Control Science and Engineering
Dalian University of Technology
Dalian, China
hanzhongyang@dlut.edu.cn

Jun Zhao
School of Control Science and Engineering
Dalian University of Technology
Dalian, China
zhaoj@dlut.edu.cn

Wei Wang
School of Control Science and Engineering
Dalian University of Technology
Dalian, China
wangwei@dlut.edu.cn

Feng Jin
School of Control Science and Engineering
Dalian University of Technology
Dalian, China
jinfeng_1126@126.com

***Abstract*—Named Entity Recognition (NER) is an important branch of Natural Language Processing (NLP). Among the existed NER methods, one of the most advanced and commonly deployed approach is the Long Short Term Memory with a Conditional Random Field layer (LSTM-CRF). However, this supervised method generally requires a large number of labeled corpuses, which is very limited regarding the texts in drug patent of this study. Bearing this in mind, a word similarity feature-based semi-supervised NER approach is proposed in this study. The feature of word similarity with regard to various types of entities are firstly extracted from word embedding to form similarity constraint. Then they are combined with the features computed by supervised LSTM. Finally, the tagged results are obtained through the CRF layer. By introducing the similarity feature of word embedding to LSTM-CRF model, the proposed method can greatly reduce the untagged cases in a large amount of similar entities. Experimental studies demonstrated that the proposed method performs obvious advantages in both the accuracy and comprehensiveness when compared with the traditional baseline model and other semi-supervised methods.**

Keywords—Named entity recognition, semi-supervised learning, word similarity, drug patents

I. Introduction

Aims at locating and classifying named entity into pre-defined categories of a specific domain, Named Entity Recognition (NER) is a pivotal subtask of Natural Language Processing (NLP). After been firstly proposed at the MUC-6 conference (the sixth in a series of Message Understanding Conferences) [1], the NER now plays a vital role in entity relationship extraction, text information extraction, machine translation and other advanced tasks. In recent years, with the rapid development of the pharmaceutical industry, a wealth of research results have emerged and many researchers began to explore the application of NER on literatures in this field. One of its greatest challenge is that these chemical named entities are professional and difficult to understand than common written words. Besides, the words should be identified not only as chemical entity, but also into detailed specific type, such as family, systematic, etc., so as to provide comprehensive information for drug discovery and development.

Initially, the NER models for texts relating to chemistry and pharmaceutics were mostly based on human experience to construct rules for tagging [2]. However, these approaches are extremely dependent on domain expertise, as well as being time-consuming along with high costs. At present, with the increasing popularity of deep learning for machine learning, researchers have proposed various neural network architectures for NER [3][4]. In 2011, R. Collobert et al. proposed a convolution neural network (CNN)-based method, which requires far less prior knowledge and expanded the application scope of the model [5]. In 2014, Kaisheng et al. improved the performance of sequence tagging by combining RNN and CRF without using hand-engineered features or data preprocessing [6]. In 2015, Z. Huang et al. compared various LSTM models and demonstrated the superiority of the BiLSTN-CRF model [7]. Due to the consideration of the dependence among input texts as well as between adjacent tags, the BiLSTM-CRF has currently achieved the state-of-the-art performances for NER task [8][9]. Although these supervised learning models are theoretically suitable for all kinds of sequence tagging tasks, the application case on professional fields is still limited because of the lack of labeled corpus. In such a case, introducing unsupervised technique is necessary for effectively solve this problem. Some literatures have already made such combinations. For example, Turian et al. utilized unsupervised word embedding as additional features of semi-supervised learning so as to improve the tagging accuracy [10]. Common semi-supervised NER models also utilize other unsupervised mapping results as additional features, e.g., word embedding in high-dimensional discrete processing [11], context embedding in language model (LM) training [12], etc. All of the above-mentioned semi-supervised learning methods simply regarded the features of a single input word as additional features of the input by unsupervised learning. While considering the significance of similarity feature among the same entity type for the professional field in this study, the features should be extracted in the perspective of the whole corpus.

In order to overcome the limitations of the existed methods, we propose a word similarity feature-based approach for NER in this study. By incorporating supervised BiLSTM-CRF model with Word Similarity (WS) features obtained from

This work is supported by the Fundamental Research Funds for the Central Universities of China (No. DUT18TD07).

unsupervised word embedding model, a semi-supervised architecture, i.e., WS-BiLSTM-CRF is constructed so as to take the advantage of both two methods. In detail, the word similarity features with regard to various types of entities are extracted to formulate similarity constraint at first. Then they are combined with the features computed by supervised BiLSTM in the form of vectors. Finally, the tagged results are obtained by CRF layer. In this way, the proposed method sufficiently utilizes the features contained in the word. And it remarkably reduces the untagged cases as well as further improve the accuracy and comprehensiveness of entity recognition. Experimental studies demonstrated that, by introducing the similarity feature of word embedding to BiLSTM-CRF model, the proposed method performs obvious superiorities for NER comparing with the traditional baseline model and other commonly deployed semi-supervised methods.

The rest of this paper is organized as follows. Section Ⅱ introduces relevant knowledge of the baseline model adopted in this paper. Section Ⅲ describes the proposed feature-based semi-supervised method for NER. Section Ⅳ reports the experimental results involving detailed settings and comparative studies. The advantages and possible future topic are summarized in the end as Section Ⅴ.

II. Preliminaries

A. Word Embedding and Silmilarity

The distributional hypothesis for words was first proposed by Harris in 1954 [13]. The core of this theory is the modeling of the relationship between target word and its context. Currently, the distributional representation of words is widely deployed in the field of NLP, especially the representation based on neural network known as Word Embedding. Some previous studies have proved that this method can automatically learn and express complex context relations from unlabeled corpus in large scale [14][15]. Since the essence of word embedding training is to describe the word with its context, the closer the corresponding word embeddings located in spatial distribution, the more similar the context of two words is. Besides, the two words have similar contexts means they have high semantic similarity. In order to express the similarity feature between words, Word Similarity (WS) is introduced in this study which is defined as the distance of word vectors in space. Obviously, WS is related to embedding distribution. In other words, the denser the word embedding distribution is, the greater the word similarity exhibits.

Regarding a professional field, such as pharmaceutical in this study, the vocabulary in context of an entity towards the same type typically exhibits some similarity. As such, utilizing this property will effectively improve the accuracy and comprehensiveness of NER.

B. Baseline

We selected BiLSTM-CRF model as the baseline model for its outstanding performance. Given a input sentence, the baseline model aims at predicting a tag for each word within. It starts with initializing each word in the input sentence by pre-trained word embedding. Then the features are automatically extracted and the confidence score for each word are obtained by the following BiLSTM network. Finally, the tag for each word in the input sentence is predicted by jointing the confidence scores and the transition scores learned from a CRF layer. The architecture of the model can be depicted as Fig. 1. In addition, brief descriptions of BiLSTM and CRF are provided as follows.

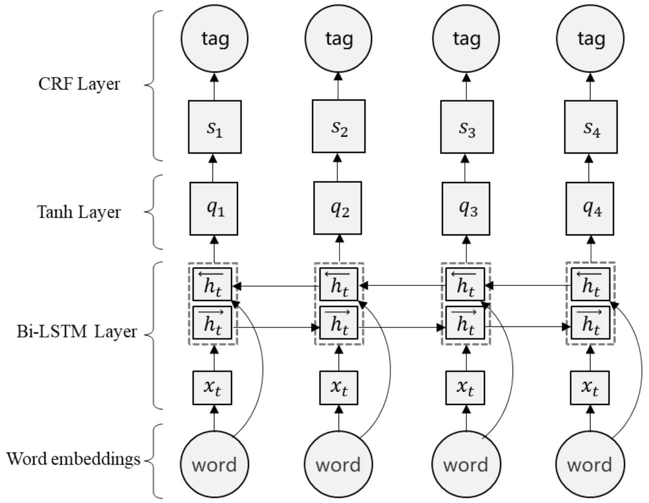

Fig. 1. The architecture of BiLSTM-CRF model.

The detailed implementation method [3] can be described as (1)-(5). For a given sentence $\mathbf{X}=(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_t,...,\mathbf{x}_n)$ containing n words, two LSTMs (forward and backward) generate representations of the left context and right context ($\overrightarrow{\mathbf{h}_t}$ and $\overleftarrow{\mathbf{h}_t}$), respectively. The final representation of a word is obtained by concatenating the two representations, $\mathbf{h}_t=\left[\overrightarrow{\mathbf{h}_t},\overleftarrow{\mathbf{h}_t}\right]$ [16]. By means of such bidirectional learning, all possible representations of a word in the context can be included, which endows BiLSTM outstanding performance of sequence tagging with numerous labeled corpus.

$$\mathbf{i}_t=\sigma\left(\mathbf{W}_{xi}\mathbf{x}_t+\mathbf{W}_{hi}\mathbf{h}_{t-1}+\mathbf{W}_{ci}\mathbf{c}_{t-1}+\mathbf{b}_i\right) \tag{1}$$

$$\mathbf{f}_t=1-\mathbf{i}_t \tag{2}$$

$$\mathbf{c}_t=\mathbf{f}_t\times\mathbf{c}_{t-1}+\mathbf{i}_t\times\tanh\left(\mathbf{W}_{xc}\mathbf{x}_t+\mathbf{W}_{hc}\mathbf{h}_{t-1}+\mathbf{b}_c\right) \tag{3}$$

$$\mathbf{o}_t=\sigma\left(\mathbf{W}_{xo}\mathbf{x}_t+\mathbf{W}_{ho}\mathbf{h}_{t-i}+\mathbf{W}_{co}\mathbf{c}_{t-i}+\mathbf{b}_o\right) \tag{4}$$

$$\mathbf{h}_t=\mathbf{o}_t\times\tanh\left(\mathbf{c}_t\right) \tag{5}$$

where σ is the element-wise sigmoid function. $\mathbf{i}_t$ and $\mathbf{f}_t$ represent the “input gate” and “forget gate”, respectively, deciding what information will be updated and what will be discarded from the cell state. $\mathbf{c}_t$ denotes current cell state. $\mathbf{o}_t$ represents the "output gate" deciding to output which parts of the cell state[17].

Considering the dependencies across output tags, the CRF layer is added to predict the best tag sequence among all possible tag sequences. As a result, the score of an input sentence $\mathbf{X}=(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_t,...,\mathbf{x}_n)$ along with a sequence of predicted tags $y=(y_1,y_2,\ldots,y_t,\ldots,x_n)$ is then given by the sum of transition scores and network scores:

$$s(\mathbf{X},y)=\sum_{i=0}^{n}T_{y_i,y_{i+1}}+\sum_{i=1}^{n}S_{i,y_i} \tag{6}$$

where S_{i,y_i} corresponds to the score of the tag y_i of the i^{th} input word in the sentence. $T_{y_i,y_{i+1}}$ represents the score of a transition from tag y_i to y_{i+1}.

The conditional probability $p(y\,|\,\mathbf{X})$ for one predicted sequence is obtained by a Softmax function over all possible tag sequences. The goal of the model is to maximize the logarithmic probability of the correct tag sequence, so that the objective function is formulated as

$$\log\left(p(y\,|\,\mathbf{X})\right)=s(\mathbf{X},y)-\log\left(\sum\nolimits_{\tilde{y}\in Y_{\mathbf{X}}}\exp\left(s(\mathbf{X},y)\right)\right) \tag{7}$$

where $Y_{\mathbf{X}}$ is the set of all possible tag sequences for input sentence $\mathbf{X}$. As a result, a valid sequence of output tags are generated. In decoding process, the tag sequence with the maximum score can be obtained by (8) which can be computed by implementing dynamic programming technique.

$$y^{*}=\underset{\tilde{y}\in Y_{\mathbf{X}}}{\arg\max}\, s(\mathbf{X},\tilde{y}) \tag{8}$$

III. Word Similarity Feature-based Semi-supervised Model

Being similar to the conventional machine learning-based NER models, the BiLSTM-CRF mentioned above is also a supervised learning method, of which the results tend to be unsatisfied when considering limited labeled corpus. In the NER task for a certain professional field, entities of the same type behave high word similarity, but only some of them can be currently tagged. For solving such problems, we designed a feature-based semi-supervised NER model integrating word similarity features into the BiLSTM-CRF model, in which the untagged parts can be identified by considering tagged parts and similarity constraint between these two parts.

The architecture of our proposed model is depicted in Fig. 2. It is mainly composed of three parts: BiLSTM layer to capture the context information of long-range dependencies, Word Similarity (WS) layer to extract the similarity features of word embedding, and CRF layer to extract the transfer features between tags and combine the state features acquired from BiLSTM and WS. The model is abbreviated as WS-BiLSTM-CRF in the following context.

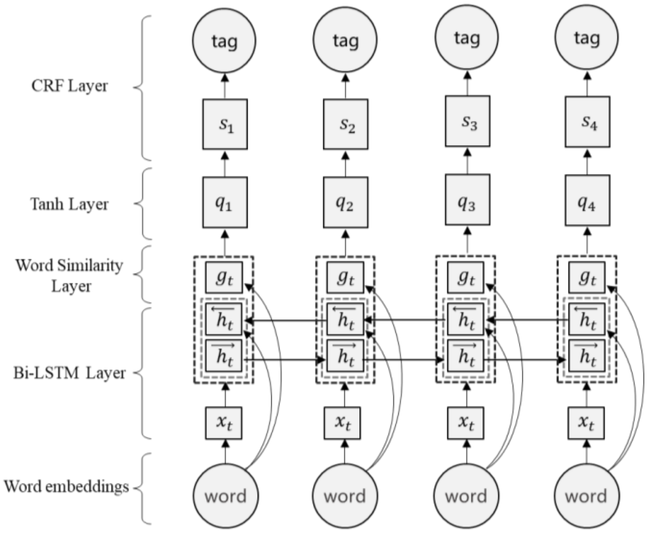

Fig. 2. The architecture of the proposed WS-BiLSTM-CRF model.

A. Word Similarity Feature Extraction

Based on the nature that the unsupervised-based word embeddings for words in labeled corpus are spatially centered, we generate a reference vector for every label by integrating word embeddings of all the words belonging to this label in the corpus. The calculation of the reference vector is inspired by the way of computing prototypes in Fuzzy C-Means clustering algorithm. Then the similarity can be quantified as the distance between the word embedding of the input word and the reference vector of each label.

In this study, we investigate the following four distance measurements, i.e., Euclidean distance, normalized Euclidean distance, Manhattan distance and Cosine distance, which are defined as follows

$$d_e(\mathbf{x}_t,\mathbf{v}_k)=\sqrt{\sum_{i=1}^{n}(x_{ti}-v_{ki})^2} \tag{9}$$

$$d_s(\mathbf{x}_t,\mathbf{v}_k)=\sqrt{\sum_{i=1}^{n}\left(\frac{x_{ti}-v_{ki}}{s_i}\right)^2} \tag{10}$$

$$d_m(\mathbf{x}_t,\mathbf{v}_k)=\sum_{i=1}^{n}|x_{ti}-v_{ki}| \tag{11}$$

$$d_c\left(\mathbf{x}_t,\mathbf{v}_k\right)=\frac{\sum_{i=1}^{n}\left(x_{ti}\times v_{ki}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{ti}\right)^2}\times\sqrt{\sum_{i=1}^{n}\left(v_{ki}\right)^2}} \tag{12}$$

where $\mathbf{x}_t$ represents the current input word embedding, and $\mathbf{v}_k$ represents the reference vector of the k^{th} label.

From the above definitions, one can easily conclude that the value of Cosine distance tends to be high when the input word embedding $\mathbf{x}_t$ and the reference vector $\mathbf{v}_k$ are close, i.e., exhibits high similarity. As such, we directly normalize the value of Cosine distance to obtain the possibility score of the input word $\mathbf{x}_t$ that it belongs to different labels whose reference vectors are $\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_k,...,\mathbf{v}_l$, respectively. While the Manhattan distance, Euclidean distance and normalized Euclidean distance behaves in an opposite direction. Thus, we regard the reciprocal of their values as $s\left(\mathbf{x}_t,\mathbf{v}_k\right)$ and then take the normalized results as the final scores (as (13)), so that their scores will be also positively correlated with the similarity.

$$w_{t,k}=\frac{\exp\left(s\left(\mathbf{x}_t,\mathbf{v}_k\right)\right)}{\sum_{k=1}^{l}\exp\left(s\left(\mathbf{x}_t,\mathbf{v}_k\right)\right)} \tag{13}$$

where l is the number of labels. The similarity feature for each input word is extracted as a weighted sum of each reference vector.

$$\mathbf{g}_t=\sum_{j=i}^{k}w_{t,k}\mathbf{v}_k \tag{14}$$

B. Word Similarity Layer

We introduce the Word Similarity layer to combine the word similarity features into the baseline model, which is concatenated after the hidden layer output of BiLSTM as a vector as (15) so as to predict the confidence score that each input corresponds to different labels.

$$\mathbf{f}_t=\left[\mathbf{h}_t,\mathbf{g}_t\right] \tag{15}$$

Similar to the traditional BiLSTM-CRF model, the score of a predicted tag sequence for the input sentence is then computed as the sum of transition scores and confidence scores by CRF layer. A Softmax function is deployed to yield the conditional probability of the path. And the objective of the model is to maximize the log-probability of the correct tag sequence.

To present the overall computing procedure of the proposed WS-BiLSTM-CRF model, the pseudocode is given as Algorithm 1.

Algorithm 1 WS-BiLSTM-CRF

```
**Initialize**: $word_t\left(t=1,2,\ldots,n\right)\in sentence$
        define function $g\left(\mathbf{x}_t,\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_l\right)$, $score\left(\mathbf{f}_t\right)$
**Parameters**: $c\in N^+, m>0, \varepsilon>0$
Represent words in sentence by vector:
    $\mathbf{X}=\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_t,\ldots,\mathbf{x}_n\right)\leftarrow word_1,word_2,\ldots,word_t,\ldots,word_n$
Obtain the hidden layer output by BiLSTM model:
    $\mathbf{h}_t\leftarrow BiLSTM\left(\mathbf{X}\right)$
Compute the reference vector of each tag：
  **Repeat:**
    Randomly initialize membership matrix $\mathbf{U}^{(0)}$
    Initialize: $word_j\left(j=1,2,\ldots,n\right)\in all\ words\ of\ one\ label$
    $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_j,\ldots,\mathbf{x}_n\leftarrow word_1,word_2,\ldots,word_j,\ldots,word_n$
    **Do:**
      $k=0$
      Update reference vectors $\mathbf{c}^{(k)}$ by: $\mathbf{c}_i^{(k)}=\frac{\sum_{j=1}^{n}\left[u_{i,j}^{(k)}\right]^m\mathbf{x}_j}{\sum_{j=1}^{n}\left[u_{i,j}^{(k)}\right]^m}$
      Compute value function $Q^{(k)}$ by:
          $Q^{(k)}=\sum_{i=1}^{c}\sum_{j}^{n}\left[u_{i,j}^{(k)}\right]^m\left\|\mathbf{x}_j-\mathbf{c}_i^{(k)}\right\|^2$
      Update membership matrix $\mathbf{U}^{(k)}$ by:
          $u_{i.j}^{(k)}=\frac{1}{\sum_{k=1}^{c}\left(\frac{\mathbf{x}_j-\mathbf{c}_i^{(k)}}{\mathbf{x}_j-\mathbf{c}_k^{(k)}}\right)^{2/(m-1)}}$
      $k=k+1$
    **While** $\left(\left|Q^{(k)}-Q^{(k-1)}\right|<\varepsilon\right)$
    Compute $s\left(\mathbf{x}_t,\mathbf{c}_i\right)$ by pre-determined measurement ($d_c$ or the
        reciprocal value of $d_e$, $d_s$ or $d_m$)
    Obtain final reference vector: $\mathbf{v}_k=\mathbf{c}_i^*=\arg\max_{0\le i\le c,i\in N}s\left(\mathbf{x}_t,\mathbf{c}_i\right)$,
$\mathbf{g}_t\leftarrow g\left(\mathbf{x}_t,\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_k,\ldots,\mathbf{v}_l\right)$
$\mathbf{f}_t=\left[\mathbf{h}_t,\mathbf{g}_t\right]$
$S_{i,y_i}\leftarrow score\left(\mathbf{f}_t\right)$
Obtain the score of a predicted tag sequence by CRF layer:
    $s\left(\mathbf{X},y\right)\leftarrow CRF\left(T_{y_i,y_{i+1}},S_{i,y_i}\right)$
Optimal tag sequence: $y^*=\arg\max_{\tilde{y}\in Y_{\mathbf{X}}}s\left(\mathbf{X},\tilde{y}\right)$,
```

IV. Experiment and Analysis

A. Experimental Settings

1) Pretraining: Deep learning-based tools, such as word2vec [14] and GloVe [15], are now widely used in NLP field for acquring word distributed representation. In order to accurately obtain semantic characteristics of chemical patent texts from a large amount of corpus being similar to them so as to achieve satisfied word embedding representation, we select 3000 unlabeled chemical patents in the “US Patent and Trademark Office” website and the patent texts contained in a standard labeled corpus (CHEMDNER-patents CEMP corpus of Bio-Creative-Ⅴ task) including 7000 abstracts as the training set. Then, word embedding is obtained by the skip-gram model of open source word2vec tool, which is capable of effectively learning features for words that do not appear frequently.

2) Data Sets: Two standard corpora are selected as the data sets of WS-BiLSTM-CRF model: one is from a large chemical patent corpus produced by Saber a. Akhondi et al. [18]. Another is CHEMDNER-patents CEMP corpus of Bio-Creative-V. A total of 7000 abstracts are provided in the sample set. The corpus contains plain-text files with three indexes (Patent identifier, Title of the patent and Abstract of the patent), and chemical entity annotated text files with six indexes (Patent identifier, Type of text from which the annotation was derived, Start offset, End offset, Text string of the entity mention, Type of chemical entity mention). The corpus contains seven different types of named entities: ABBREVIATION, FAMILY, FORMULA, IDENTIFIERS, MULTIPLE, SYSTEMATIC, TRIVIAL.

3) Tagging Scheme: A single named entity could cover several words in an input sentence. Thus, the entity in this study is represented according to the BIO (Beginning, Inside, Outside) format, i.e., if a word is the beginning of an entity, it is labeled as "B-label", and inside as "I-label". If it does not belong to an entity, that is, outside an entity, it is labeled as "O". With the tag BI, we get the chunk that represents an entity name.

The experiments are divided into two groups. One is comparing the performances of four types of distance for determining the approach of quantifying similarity, and another is compared with the baseline model and the state-of-the-art semi-supervised learning model respectively to verify the effectiveness of the proposed model.

B. Comparative study for word similarity representation

Four types of distance for quantifying similarity, i.e., Euclidean, Normalized Euclidean, Manhattan and Cosine, are selected and compared in order to choose the best computing method. We have carried out an experimental comparison on the CHEMDNER-patents CEMP corpus, of which the results are given in TABLE I. .

TABLE I. TAGGING RESULT USING DIFFERENT SIMILARITY FUNCTIONS ON THE CHEMDNER-PATENTS CEMP CORPUS

Measurement	Precision	Recall	F1 score
Euclidean distance	84.67	88.39	86.49
normalized Euclidean distance	84.86	88.35	86.57
Manhattan distance	84.49	88.16	86.28
Cosine distance	**86.03**	**88.72**	**87.35**

It is demonstrated that Precision and Recall of the model with the word similarity quantified by cosine distance are better than other algorithms, and the F1 value is the highest (87.35%). Although there is little difference between the four methods in terms of single index, we choose cosine similarity as the word similarity quantization method considering the comprehensive performance and the physics meaning of these four methods.

C. Comparative study for NER

1) Comparing with Baseline

In order to study the influence of similarity features on the performance of the model, we test this semi-supervised learning method and the baseline model for NER on the CHEMDNER-patents CEMP corpus and another real-world drug patent.

TABLE II. TAGGING RESULTS OF BASELINE AND PROPOSED METHODS ON THE CHEMDNER-PATENTS CEMP CORPUS

Model	Precision	Recall	F1 score
Baseline (BiLSTM-CRF)	83.03	87.27	85.09
WS-BiLSTM-CRF	**86.03**	**88.72**	**87.35**

TABLE III. TAGGING RESULTS OF BASELINE AND PROPOSED METHODS ON A DRUG PATENT

Model	Precision	Recall	F1 score
Baseline (BiLSTM-CRF)	78.88	79.73	79.30
WS-BiLSTM-CRF	**84.98**	**86.57**	**85.76**

The experimental results are presented as TABLE II. and TABLE III. . With regard to the standard CEMP corpus, the Precision and Recall of our model are substantially improved, as well as the comprehensive evaluation index F1 reaching 87.35%, i.e., an increase of 2.26% comparing to the baseline model. As for the tagging of the chemical entities in practical patent texts, the model incorporating the similarity features of the word also behaves superiority, with the F1 score of 85.76%, which is almost 7% higher than the baseline model. It is obvious that the model proposed in this paper exhibits outstanding performance on both accuracy and comprehensiveness.

Although the BiLSTM-CRF model performs well in most NER tasks, the context information it learns in the above experiments is very limited due to the few number of the corpus. Owing to the fact that more unlabeled corpus can be obtained for training, word embedding based on unsupervised learning contains much more context information than that learned by supervised models. Therefore, the proposed semi-supervised NER model consistently provide better results.

2) Comparing with other commonly deployed semi-surpervised methods

In order to further verify the effectiveness of the proposed approach, we compare it with other commonly deployed semi-supervised learning NER model. One takes word embedding as additional features and another with Language Model (LM) embedding are selected to conduct on the CHEMDNER-patents CEMP corpus and the new drug patent, respectively.

The statistics of the results obtained by these three models are presented in TABLE IV. and TABLE V. . Both the accuracy and comprehensiveness of chemical NER performs better than the advanced semi-supervised learning model, especially for the application on patent texts.

TABLE IV. TAGGING RESULTS OF DIFFERENNT SEMI-SUPERVISED METHODS ON THE CHEMDNER-PATENTS CEMP CORPUS

Model	Precision	Recall	F1 score
BiLSTM-CRF(+emb)	83.59	86.46	85.18
BiLSTM-CRF(+LM)	84.57	86.55	85.55
WS-BiLSTM-CRF	**86.03**	**88.72**	**87.35**

TABLE V. TAGGING RESULTS OF DIFFERENNT SEMI-SUPERVISED METHODS ON A DRUG PATENT

Model	Precision	Recall	F1 score
BiLSTM-CRF(+emb)	81.43	82.87	82.14
BiLSTM-CRF(+LM)	81.94	82.53	82.23
WS-BiLSTM-CRF	**84.98**	**86.57**	**85.76**

Although the traditional semi-supervised learning models are capable of processing unlabeled corpus, the relationship between words in these methods are only reflected by the relative position in vector space. The insufficient utilization of similarity feature of these models results in weak correlation between words along with unsatisfied performance on solving some NER problems. As for professional literatures (such as drug patents) with highly specialized entities and similar contexts, extraction and deployment of similarity features in word embedding can effectively improve accuracy of NER. Instead of implicitly expressing the relationship among word embeddings by the position of input words in the vector space, we explicitly quantify it to provide reliable guidance for the whole proposed model. In such a way that directly integrating the unsupervised feature into the supervised baseline model, the influence of the similarity features on the tagging process is strengthened. As a result, the proposed model achieves outstanding performance.

V. CONCLUSIONS

In this paper, we propose a semi-supervised approach for the chemical NER task. The feature of word similarity are extracted from word embedding to form similarity constraint at first, and then combined with the features learned by supervised LSTM. The final tagged results are obtained through the CRF layer. The experimental results demonstrate that: (a) introducing feature of word embedding to use the similarity information between words can successfully reduce the untagged case of similar entities; (b) the proposed WS-BiLSTM-CRF model intentionally takes the characteristics of word embedding into account, and it outperforms other semi-supervised learning methods. Because of these two advantages, it can achieve the optimal performance on both chemical NER task of CEMP corpus and real-world patent (87.35% and 85.76% respectively in F1 score).

Our WS-BiLSTM-CRF method is particularly applicable to professional fields, such as the drug patent in this paper, as well as chemical literature, legal document, etc. While considering the generalization of the method, research on NER for the general field, such as press release and novel, is a possible future topic deserving further study.

REFERENCES

[1] R. Grishman and B. Sundheim, "Message understanding conference-6: A brief history,", 1996.

[2] T. Rocktäschel, M. Weidlich and U. Leser, "ChemSpot: a hybrid system for chemical named entity recognition," Bioinformatics, 28, pp. 1633-1640, 2012.

[3] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," arXiv preprint arXiv:1603.01360, 2016.

[4] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional lstm-cnns-crf," arXiv preprint arXiv:1603.01354, 2016.

[5] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," J. Mach. Learn. Res., 12, pp. 2493-2537, 2011.

[6] K. Yao, B. Peng, G. Zweig, D. Yu, X. Li, and F. Gao, "Recurrent Conditional Random Field for Language Understanding,", 2014.

[7] Z. Huang, W. Xu and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," arXiv preprint arXiv:1508.01991, 2015.

[8] L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin, and J. Wang, "An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition," Bioinformatics, 34, pp. 1381-1388, 2017.

[9] D. Zeng, C. Sun, L. Lin, and B. Liu, "LSTM-CRF for drug-named entity recognition," Entropy-Switz., 19, p. 283, 2017.

[10] J. Turian, L. Ratinov and Y. Bengio, "Word representations: a simple and general method for semi-supervised learning,", 2010, pp. 384-394.

[11] J. Guo, W. Che, H. Wang, and T. Liu, "Revisiting embedding features for simple semi-supervised learning,", 2014, pp. 110-120.

[12] M. E. Peters, W. Ammar, C. Bhagavatula, and R. Power, "Semi-supervised sequence tagging with bidirectional language models," arXiv preprint arXiv:1705.00108, 2017.

[13] Z. S. Harris, "Distributional structure," Word, 10, pp. 146-162, 1954

[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality,", 2013, pp. 3111-3119.

[15] J. Pennington, R. Socher and C. Manning, "Glove: Global vectors for word representation,", 2014, pp. 1532-1543.

[16] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," Neural Networks, 18, pp. 602-610, 2005.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., 9, pp. 1735-1780, 1997.

[18] A. Akhondi, A. G. Klenner, C. Tyrchan, A. K. Manchala, K. Boppana, D. Lowe, M. Zimmermann, S. A. Jagarlapudi, R. Sayle, and J. A. Kors, "Annotated chemical patent corpus: a gold standard for text mining," PLoS One, 9, p. e107477, 2014.